

## ABSTRACT

A computer method for representing a natural-language document in a vector form suitable for text manipulation operations is disclosed. The method involves determining (a) for each of a plurality of terms selected from one of (i) non-generic words in the document, (ii) proximately arranged word groups in the document, and (iii) a combination of (i) and (ii), a selectivity value of the term related to the frequency of occurrence of that term in a library of texts in one field, relative to the frequency of occurrence of the same term in one or more other libraries of texts in one or more other fields, respectively. The document is represented as a vector of terms, where the coefficient assigned to each term includes a function of the selectivity value determined for that term, and optionally related to the inverse document frequency of that word in one or more libraries of texts. Also disclosed are a computer-readable code for carrying out the method, a computer system that employs the code, and a vector generated by the method.

15